Objective Assessment of Karate Kata Performance through Synchronized Multi-View Video Analysis and Deep Learning

[1]Bevin Benet, [2]Sreelekshman Sreekumar, [3]Vishnu Mohan C
[1] Department of Computer Science, Sacred Heart College (Autonomous), Thevara
[2] Department of Computer Science, Sacred Heart College (Autonomous), Thevara
[3] Department of Computer Science, Sacred Heart College (Autonomous), Thevara
[1]bevinbenet@gmail.com[2] sreelekshmanchunakara@gmail.com [3]vishnumohan@shcollege.ac.in

Abstract*— Karate kata play a central role in grading, competition, and the development of core psychomotor skills. Traditional human-judged evaluations can be susceptible to subjectivity and inter-rater variability. This challenge aligns closely with the broader field of Action Quality Assessment (AQA), which focuses on evaluating the quality of human actions in a way that mimics expert judgment. In this work, we introduce a novel regression pipeline that integrates synchronized dual-view video capture (frontal and lateral) with markerless pose estimation and Long Short-Term Memory (LSTM)–based regression to generate objective, continuous quality scores for entire kata performances. As a starting point, we have chosen the first kata of the Shōrin-ryū style of karate named Taikyoku Shodan. A total of 50 kata executions were recorded simultaneously from two calibrated camera perspectives and annotated by an expert judge. MediaPipe was used to extract and normalize 2D skeletal keypoints, filtering out the frames from which angles cannot be extracted correctly using the face orientation. Frame-wise joint angles were then computed and assembled into time-series feature vectors.To address data scarcity, we augmented the sequences using temporal shifts, Gaussian noise injection, and slight amplitude scaling, expanding the dataset to 2,400 training samples. The vectors were fed into a bidirectional LSTM network that learns temporal dependencies across the full kata sequence. Our supervised regression model achieved a mean squared error (MSE) of 0.141 and a Spearman's rank correlation coefficient of 0.802 compared to expert scores, demonstrating high fidelity in replicating human judgment.

Index Terms— ***Karate, Kata, Pose Estimation, Computer Vision, Data Augmentation.***

I. INTRODUCTION

Karate, a widely practiced martial art, emphasizes precision, discipline, and control through kata — a series of choreographed techniques that reflect the core principles and philosophy of the discipline. Mastery of kata is foundational not only for progression through ranks and success in competitive environments, but also for cultivating psychomotor abilities such as balance, coordination, timing, and power generation [35,36]. Traditionally, kata performances are evaluated by panels of human judges who assess technical execution, rhythm, power, and overall quality. However, such evaluations are inherently subjective and prone to inter-rater variability. Furthermore, scaling this evaluation process to large cohorts—such as in national examinations or digital learning platforms—poses significant logistical challenges.

The advent of modern computer vision and deep learning techniques offers powerful tools for automating human movement analysis in a precise and repeatable manner. Markerless pose estimation frameworks like OpenPose and MediaPipe can extract skeletal keypoints from monocular video streams with high accuracy [16,32], enabling fine-grained spatiotemporal representation of complex human motion. When combined with deep sequence models—such as Long Short-Term Memory (LSTM)[11,12], networks—these representations can be used to capture intricate motion dynamics across time, effectively modeling the temporal continuity inherent in martial arts sequences [34].

This direction aligns closely with the broader field of Action Quality Assessment (AQA), which focuses on evaluating the quality of human actions in a manner that mimics expert judgment. AQA systems aim to assign performance scores to complex activities by analyzing visual motion patterns, often using video-based inputs. Unlike simple action classification, AQA targets skill-based evaluation, where small variations in execution can lead to significant differences in scoring. It emphasizes not only accuracy and consistency, but also interpretability and generalizability, making it a valuable framework for domains like sports, rehabilitation, and martial arts [37,38,26,29].

While previous studies have explored applications of pose estimation in tasks such as individual action recognition, multimodal interaction modeling, or real-time combat feedback [1,10,18], few systems have focused specifically on kata assessment. Even fewer aim to regress a single, objective score for an entire kata performance that mimics expert judgment. This leaves a gap in developing a unified, scalable, and unbiased scoring mechanism suitable for educational and competitive use.

To address this gap, our research presents a novel kata scoring pipeline that combines synchronized dual-view video recording (frontal and lateral) with pose estimation, joint-angle computation, and deep learning-based temporal modeling. By capturing two orthogonal views of each kata performance as illustrated in Figure 1, the system achieves a more comprehensive spatial understanding of movements. Pose landmarks extracted via MediaPipe are filtered and transformed into time-series joint angles, which are then processed through a bidirectional LSTM network to predict expert-style performance scores.

Funded by HI-BIRDS 2024-25 (Heartian Initiative for Building Innovation, Research and Development Scheme) of Sacred Heart College. Supported by Shorin Ryu Seibukan Karate Academy of India (SSKAI) in labeling the kata videos used in this study..

Figure 1: Frontal and Lateral View from Camera

As a proof of concept for this supervised regression approach, we focus our investigation on Taikyoku Shodan, the first kata of the Shōrin-ryū style of karate. This fundamental kata provides an ideal testing ground for our methodology, as it encompasses essential karate techniques while maintaining sufficient complexity to validate the effectiveness of our automated scoring system. Our regression model demonstrates the feasibility of objective kata assessment and establishes a foundation for future expansion to more complex kata forms.

This approach not only leverages the strengths of existing pose and sequence modeling frameworks but also introduces a scalable solution for kata evaluation that is reproducible, unbiased, and suitable for integration into training environments or examination systems.

Contributions of the proposed work

- We propose a dual-view, LSTM-based scoring pipeline that regresses continuous kata quality scores from pose-estimated angle sequences.
- We demonstrate that the system generalizes well to unseen kata samples, with high correlation to expert scores, suggesting its viability as a supplemental or standalone assessment tool.
- Our ablation and comparative studies highlight the superiority of the dual-view paradigm, setting a new baseline for future research in this domain and demonstrating methodological extensibility to other kata forms and martial arts styles.

The remainder of this paper is organized as follows: Section II reviews the related work; Section III details the proposed methodology; Section IV presents the experimental analysis; Section V lists the results of the research; Section VI lists the future directions.

II. Related Work

The following background reviews key developments in computer vision, temporal modelling, and multimodal systems that form the foundation of our dual-view LSTM-based kata scoring system.

2.1. Scoring Human Actions in Video-Based AQA

Across the six studies, recent advancements in Action Quality Assessment (AQA) have centered around enhancing both the interpretability and precision of automated evaluation systems through improved feature representation, fine-grained temporal modeling, and more transparent scoring mechanisms. The foundational work by Pirsiavash et al. [37] introduced a spatiotemporal pose-based regression model to assess action quality, laying the groundwork with hierarchical features and interpretable gradient-based feedback. Building upon this, Xu et al. [38] developed FineDiving—a diving-specific dataset with procedural step annotations—and a Temporal Segmentation Attention (TSA) module that leverages fine-grained segmentation and contrastive regression, significantly improving performance on temporally sensitive actions. Similarly, Jain et al. [26] proposed a Reference Guided Regression (RGR) framework using Siamese networks and deep metric learning(DML) [20] to learn similarity-based scores between expert and test performances, achieving superior results on the UNLV Diving, Vault, and MTL-AQA datasets. This is achieved through a two-phase approach: first, a Deep Metric Learning (DML) module learns similarity metrics between action sequences using a Siamese-LSTM network, producing binary similar/dissimilar labels based on ground truth score differences. Second, a Score Estimation Module utilizes this learned similarity metric to estimate a video's score based on its resemblance to an expert reference video of the same action type.

Meanwhile, efforts by Parmar and Morris [29] introduced multiple deep learning variants combining C3D, LSTM, and SVR to extract spatiotemporal features from Olympic event videos. Their models showed the advantage of combining temporal modeling with regression for improved accuracy, especially in scenarios with limited training data. Qi et al. [22] focused on robust procedural segmentation and hierarchical pose encoding through a multi-stage contrastive regression framework, validated on FineDiving-Pose and MTL-AQA datasets, enabling better handling of motion blur and sub-action separation. Finally, Zhou et al. [21] provided a large-scale survey and a unified benchmark incorporating six datasets and seven evaluation metrics. These include MAE, MSE, rMSE, and Accuracy for score prediction, establishing standardized metrics that many of the other works adopted. Together, these papers reflect a shift toward interpretable, fine-grained, and metric-sensitive AQA systems grounded in strong empirical validation and consistent evaluation protocols.

2.2. Data Augmentation Techniques in Deep Learning

Data augmentation plays a crucial role in enhancing deep learning models, especially when dealing with limited or imbalanced datasets. In the domain of image data, augmentation techniques range from basic transformations and color manipulations to advanced methods like feature space alteration, neural style transfer, and deep generative models such as Generative Adversarial Networks (GANs).

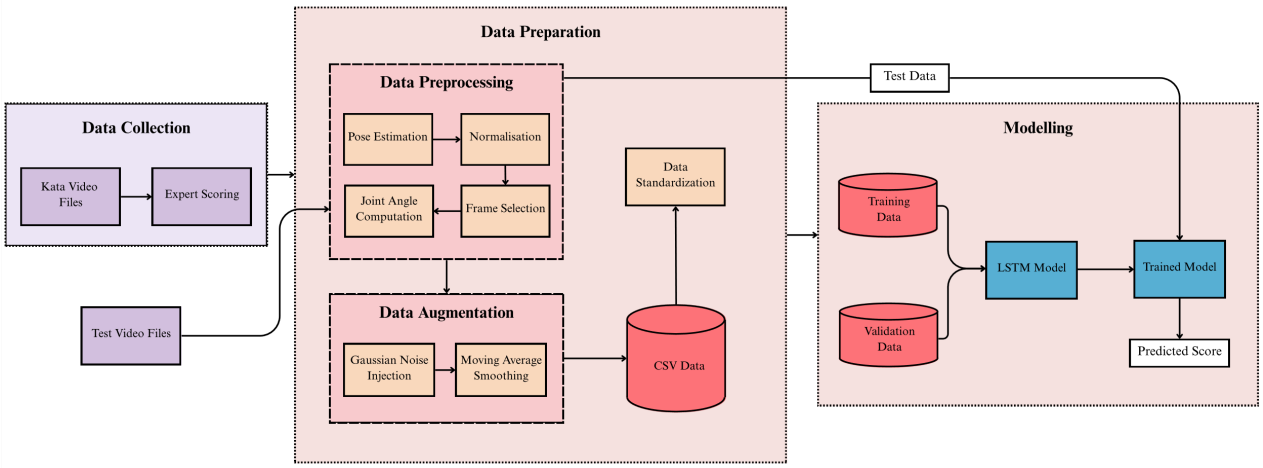

Figure 2: Methodology Workflow

These techniques not only enrich dataset diversity but also help mitigate issues like overfitting, class imbalance, and domain shifts, making them valuable across tasks such as classification, object detection, and semantic segmentation [30,31]. While traditional methods offer ease of implementation and label preservation, recent advances emphasize the utility of meta-learning frameworks and automated augmentation policies to dynamically adapt augmentation strategies based on task-specific needs.

For time series data, augmentation is equally essential due to the scarcity of labeled datasets and the susceptibility to overfitting. Approaches are broadly categorized into transformation-based, pattern-based, generative, decomposition-based, and automated methods, with empirical evaluations revealing that context and dataset characteristics heavily influence effectiveness [23-25]. Techniques such as slicing, window warping, and discriminative guided warping show consistent improvements, while automated and generative methods like TimeGAN and VAEs are gaining traction. Additionally, unified evaluation tools and libraries are being developed to guide the selection and application of appropriate augmentation methods. These advancements collectively aim to create robust, generalizable deep learning models across both visual and temporal data domains.

2.3. Motion Generation for Martial Arts Interactions

One major limitation of traditional motion generation systems is their focus on solo performances, which fail to capture the dynamic, reciprocal nature of combat sports like karate. Addressing this, Huang et al. [1] proposed MagicFight, a personalized fighting motion synthesis system that extends the Stable Diffusion framework for martial arts interactions. It introduces three modules: a reference net to ensure identity consistency, a pose guider to handle user-defined movements, and a temporal layer for smooth animation over time. These components work synergistically to simulate interactive motion sequences between two fighters, a significant improvement over previous single-person generation models [2]. By using a Unity-based dataset and a modified AnimateAnyone pipeline, MagicFight demonstrated how deep generative models could effectively handle complex motion tasks that mirror real combat behavior—an insight that is especially relevant when attempting to synthesize or analyze karate katas performed in pair training or bunkai.

2.4. Martial Arts Competition and Training Evaluation

The application of artificial intelligence and data-centric models in the field of Wushu martial arts has seen significant advancement through multiple studies. Shisen et al. [8] proposed an intelligent system for decision-making in martial arts competitions. This system uses a deep neural network to classify large datasets into a data warehouse and employs a multi-agent AI system for real-time tactical decisions. It is evaluated using metrics like accuracy, specificity, sensitivity, and Mathew's Correlation Coefficient (MCC), offering scalable and effective assistance to coaches during live competitions. Similarly, Jin [9] has developed a training evaluation system that utilizes a deep learning model with normalized input layers trained on original teaching data. This system is capable of identifying movements, offering correction suggestions, grading performance, and even providing video recordings to support student learning.

Additionally, Li et al. [27] proposed the use of MEMS sensors for capturing human posture data and predicting injury risk using a support vector regression model. The setup forms a human body sensor network and has demonstrated potential in enhancing both performance monitoring and injury prevention strategies in martial arts training.

2.5. Multimodal Feedback and Training Platforms

In addition to generation, real-time feedback systems have become a growing area of interest for performance analysis in martial arts. Echeverria and Santos [16] introduced KUMITRON, a multimodal training platform that integrates pose estimation, inertial sensors, and even drone-based imaging to analyze karate interactions. Built around the SMDD framework [19], KUMITRON captures data from both players in a sparring session and processes it to provide real-time assessments of actions and reactions [17,18]. This approach goes beyond static movement analysis to capture tactical engagement, an aspect particularly vital in kata bunkai. While KUMITRON supports two-player scenarios and includes multimodal data sources, it lacks a consistent scoring mechanism or a temporal regression approach that

quantifies kata execution quality—something our work aims to resolve.

2.6. Computer Vision for Human Action Recognition in Sports

Researches from Host and Ivašić-Kos [13] and Thomas et al. [15] provides a thorough exploration of Human Action Recognition (HAR) and computer vision applications in sports. Host and Ivašić-Kos propose a novel classification scheme for HAR that separates human movement into three categories—Individual Actions, Joint Activity, and Team Activity—based on the number of participants and intention. They proposed a new categorization inspired from the human actions categorization in [14], offering a clearer structure for organizing HAR tasks.

Expanding on this foundation, Host and Ivašić-Kos provide a comprehensive overview of the HAR pipeline, including stages such as data acquisition, preprocessing, feature extraction, and classification using both traditional (e.g., SVM, Optical Flow, HOG) and deep learning techniques (e.g., CNNs, LSTMs, Transformers). They emphasize the benefits of 3D CNNs for capturing spatio-temporal features and LSTMs for handling sequential data, while also addressing practical challenges like occlusion and rapidly changing scenes in sports. Meanwhile, Thomas et al. focus specifically on the role of computer vision in handling the fast-paced dynamics of modern sports. Their work explores crucial components such as camera calibration, ball and player tracking, and video enhancement, and they also offer a curated list of public sports datasets for training CV models. Their review highlights not just technological progress, but also the unresolved obstacles that hinder full-scale deployment in real-world settings.

2.7. Robust 3D Pose Estimation from 2D Inputs

Accurate pose estimation is the foundational part of analyzing and scoring martial arts movements. Martinez et al. [4] addressed the challenge of 3D pose reconstruction from 2D joint locations using a simple but powerful feedforward neural network. Their method, trained on Human3.6M [7] and using 2D keypoints from a stacked hourglass network [5] pre-trained on MPII [6], showed that even without sophisticated temporal modeling or motion priors, highly accurate 3D joint positions could be recovered. The key insight was that errors in 3D pose estimation largely stem from inaccuracies in the 2D input, rather than the 2D-to-3D lifting model [4]. This work is relevant in our context, as many karate kata videos are captured using single or dual camera setups, making reliable 3D estimation from 2D landmarks a prerequisite for evaluating pose quality and movement consistency.

2.8. Deep Learning Models for Kata Action Classification

Mazen Walid et al. [33] focus on karate, specifically the Taikyoku Shodan kata, developing a system that combines pose estimation via MediaPipe with both classical ML and advanced DL models to analyze sequential movements. Their custom dataset, necessitated by the absence of public resources, enabled them to compare models such as Random Forest, K-Nearest Neighbor, MLP, RNN, LSTM, CNN, and CNN-LSTM. The system provides accurate feedback, enhancing learning and minimizing injury risk. Meanwhile, Zhou [3] presents an approach to Wushu movement recognition using a refined Hidden Markov Model (HMM) trained with the Baum-Welch algorithm. The model incorporates wavelet denoising and a sliding window to process acceleration signals, outperforming CNN, SVM, and BP-based models in recognition accuracy.

Complementing these approaches, Hong-bo and Yao [28] propose a computer vision-driven model for assessing martial arts movements by extracting motion parameters and solving a non-linear optimization problem through the Levenberg-Marquardt (L-M) algorithm. This method improves judgment precision by minimizing differences between actual and modeled motion, offering higher sensitivity to the dynamic and nuanced nature of human movement. Collectively, these studies demonstrate how computational models can not only automate but also refine performance assessment in martial arts, surpassing traditional evaluation techniques through objective, repeatable, and fine-grained analysis.

2.9. Style Classification Using Temporal Deep Learning

Going further into movement quality analysis, Daker et al. focused on classifying kata styles as “good” or “bad” based on full-sequence pose data. Using a temporal modeling approach with LSTM-based classifiers, they showed that style classification performs significantly better when using entire kata sequences instead of isolated snapshots [34]. This reinforces the idea that temporal dynamics, rhythm, and consistency across the performance are crucial indicators of kata quality—something that snapshot-based systems or frame-by-frame metrics often miss. Their work supports our choice of using LSTM regressors, which naturally model the temporal evolution of the body’s pose across a kata routine, enabling a nuanced and continuous scoring output.

This body of work underscores advances in motion synthesis, pose estimation accuracy, temporal modeling, and immersive feedback. However, the need for an accurate, unbiased kata scoring system remained unaddressed. Motivated by this gap, we developed a reliable offline scoring pipeline—leveraging dual-view pose data and an LSTM-based regressor—to objectively quantify kata performance with consistency and minimal human bias. The proposed system effectively outputs a score that measures the ability of the person in doing the kata.

III. Methodology

In this section, we present the dataset as well as the present methodology used for score estimation of Taikyoku Shodan kata movements. The entire workflow of the methodology is illustrated in the Figure 2.

3.1. Dataset

The training dataset comprises joint angle trajectories extracted from pose estimations of Taikyoku Shodan kata performances. Taikyoku Shodan, the foundational kata of the Shōrin-ryū karate style, was selected as the focus of this study due to its pedagogical significance as the initial form taught to students in this discipline. This choice represents a strategic approach to model development, as establishing baseline performance on fundamental kata provides a more manageable starting point than attempting to directly address the complexities inherent in advanced kata forms. The systematic progression from basic to complex movements

aligns with established machine learning practices of validating methodologies on well-defined, foundational cases before scaling to more intricate applications.

To ensure a higher fidelity of pose estimation, each kata performance was recorded simultaneously from two orthogonal camera angles, thereby enabling more precise triangulation of joint positions and resolving ambiguities in spatial representation. Figure 3 shows a simple graphical representation of the video capturing setup used.

A total of 50 unique kata executions were captured, with each performance filmed from two distinct viewpoints, resulting in 50 pairs of synchronized video sequences. From those, 37 were used for training the model and 13 were used in the evaluation step. The kata executions were done by expert students. All recordings were recorded at 720p resolution and 30 frames per second, ensuring consistent spatial-temporal resolution suitable for downstream kinematic analysis.

To complement the motion capture data with evaluative context, each kata performance was independently assessed and scored by two expert karate masters expert in Shōrin-ryū style of karate, providing a reliable ground truth for performance quality and enabling the system to learn from domain-specific qualitative judgments. Because real-world kata competitions attract mostly high-level performers, we restricted the training set to executions that received expert scores above 5 on a 10-point scale. Very low-quality performances are easily recognizable—even by non-experts—and therefore fall outside the scope of a fine-grained scoring model aimed at differentiating strong competitors. After this filtering step, the score distribution of the training dataset was: Mean = 7.18,Min = 5.02. Max = 9.59, Skewness = 0.24.

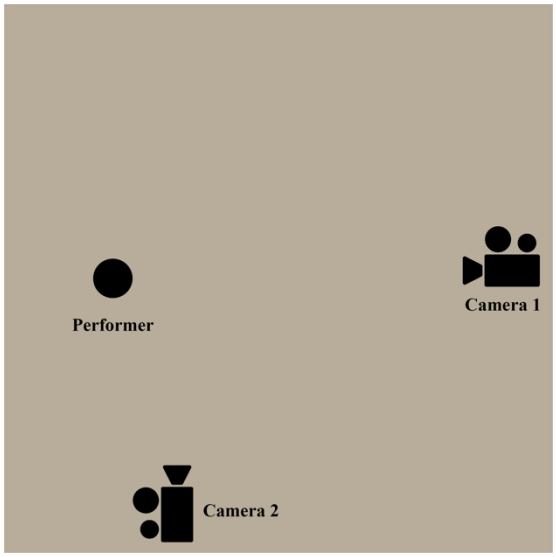

Figure 3: Top View Representation of Camera Placement

The resulting dataset is therefore slightly right-skewed, centered on the "good–excellent" range, and provides the narrow score band needed for learning subtle performance differences while reflecting the competitive environment in which the model will ultimately be deployed. This scoring mechanism introduced an additional layer of supervision, facilitating not only motion recognition but also the potential for skill-level assessment through machine learning models.

3.2. Data Preprocessing

To enable accurate and reproducible biomechanical analysis of karate kata performances, we devised a multi-stage pipeline for extracting joint angle time-series from dual-angle video recordings. Each kata sequence was captured from two synchronized viewpoints—frontal and lateral—to mitigate depth-related distortions common in monocular pose estimation.

3.2.1. Pose Estimation from Dual Views

Initially, pose landmarks were extracted from both video inputs in their entirety, without filtering for frame relevance. We employed MediaPipe Pose Estimation, a real-time model capable of generating 3D keypoints for 33 anatomical landmarks per frame as illustrated in the Figure 4, capturing the spatial configuration of the performer's body throughout the kata sequence. This extraction was performed independently on both the frontal and side-view videos, resulting in two complete sets of time-series 3D keypoints per kata.

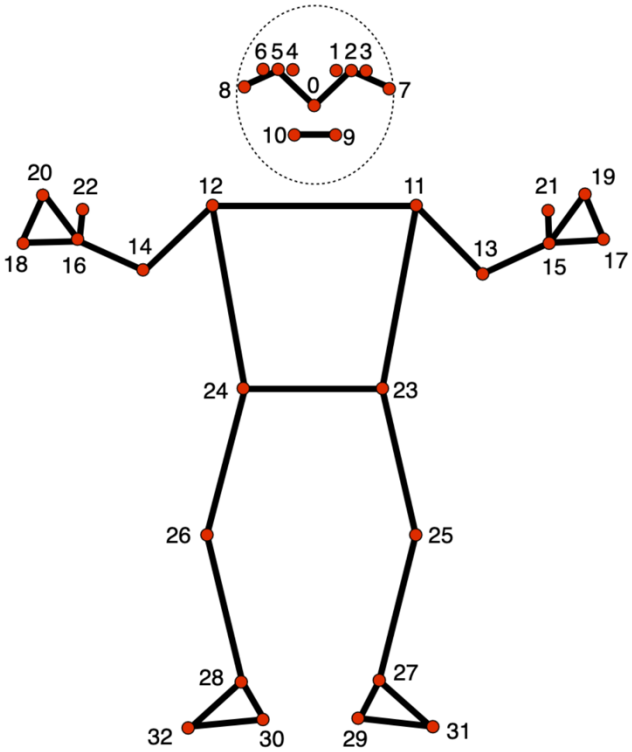

Figure 4: Mediapipe Pose Landmarks (reproduced under CC–BY 4.0 from Google AI)

The extracted pose data was then normalized by projecting the z-coordinate to a constant value, effectively transforming the 3D keypoints into a 2D representation. This projection removes depth inconsistencies, which are especially problematic when only single-camera views are used, and simplifies downstream angle computation.

3.2.2. Frame Selection Based on Face Orientation

To maximize angular measurement accuracy, we implemented a lightweight face-orientation filter that

examines each frame's yaw (left / right head rotation) and selects only those frames where the performer's face is turned away from the camera. In karate kata, limb trajectories and body planes align most naturally with the performer's gaze direction; capturing joint angles from a frontal view introduces depth ambiguity and occlusion errors. By contrast, a side-on perspective preserves the planar geometry of limb motion, yielding more reliable angle computations.

Our algorithm computes yaw by comparing the horizontal position of the nose tip to the midpoint of the eyes.
Specifically, we first determine the eyes' midpoint and measure the inter-ocular distance. We then establish a tolerance window—set at 50\% of the eye distance—around that midpoint. If the nose deviation lies within this window, the face is deemed "forward" and the frame is excluded; otherwise, it is labeled "side," indicating a side view. During kata execution, as the performer naturally pivots or shifts direction, this yaw-based test dynamically toggles which camera's frames are used: frontal frames are ignored whenever the nose is aligned within the tolerance zone, and side-view frames are ingested whenever the nose deviates beyond it.

3.2.3. Joint Angle Computation

Following frame selection, a custom angle computation module was employed to extract biomechanically relevant features from the pose landmarks. Using the reduced 2D

keypoints, the system calculates angular values at critical joints—specifically the elbows, shoulders, hips, and knees—based on the geometric relationships among adjacent landmark triplets. Each joint angle is computed using three specific landmarks as detailed in Table 1, where the middle landmark serves as the vertex point and the angle is calculated between vectors formed by the vertex to the two adjacent landmarks.

The calculation employs vector algebra and trigonometric functions to determine the internal angle at the middle joint. For each joint, the algorithm creates vectors from the vertex landmark (Point B) to the two adjacent landmarks (Points A and C), computes their dot product and magnitudes, and applies the inverse cosine function to derive the angular measurement in degrees. This approach ensures consistent and anatomically meaningful angle extraction across all frames.

Joint Angle	Point A	Vertex (Point B)	Point C
Left Elbow	12	14	16
Right Elbow	11	13	15
Left Shoulder	14	12	15
Right Shoulder	13	11	23
Left Hip	12	24	26
Right Hip	11	23	25
Left Knee	24	26	28
Right Knee	23	25	27

Table 1: Landmark Triplets for Joint Angle Calculation

The final output of this process is a frame-wise time series of joint angles, organized with reference to the original frame numbers and serialized for use in further analysis. The structure of the joint angle dataset is shown in Table 2.

The joint angle θ at vertex point B is computed using the dot product formula between two vectors formed by the landmark triplet:

$$\theta = \cos^{-1}\left(\frac{\overrightarrow{BA} \cdot \overrightarrow{BC}}{\left|\overrightarrow{BA}\right| \cdot \left|\overrightarrow{BC}\right|}\right)$$

where:

$$\overrightarrow{BA} = \left(A_x - B_x, A_y - B_y, A_z - B_z\right)$$

$$\overrightarrow{BC} = \left(C_x - B_x, C_y - B_y, C_z - B_z\right)$$

$$\left|\overrightarrow{BA}\right| = \sqrt{(A_x - B_x)^2 + \left(A_y - B_y\right)^2 + (A_z - B_z)^2}$$

$$\left|\overrightarrow{BC}\right| = \sqrt{(C_x - B_x)^2 + \left(C_y - B_y\right)^2 + (C_z - B_z)^2}$$

Column Data	Type
frame number	int64
left elbow	float64
Right elbow	float64
left shoulder	float64
Right shoulder	float64
left hip	float64
right hip	float64
left knee	float64
right knee	float64

Table 2: Structure of joint angle dataset

3.3. Data Augmentation

Due to the relatively limited number of annotated kata sequences available for training, an advanced data augmentation framework was designed to synthetically increase both the size and diversity of the dataset. This process enhances the model's capacity to generalize across performers and execution styles. Specifically, from the original pool of 50 recorded kata executions, 37 videos were selected for training. These 37 samples were programmatically augmented to produce a total of 2,400 samples, significantly expanding the dataset.

Subsequently, a moving average smoothing filter—with a randomly selected window size—is applied to the jittered data. This attenuates abrupt transitions and creates more

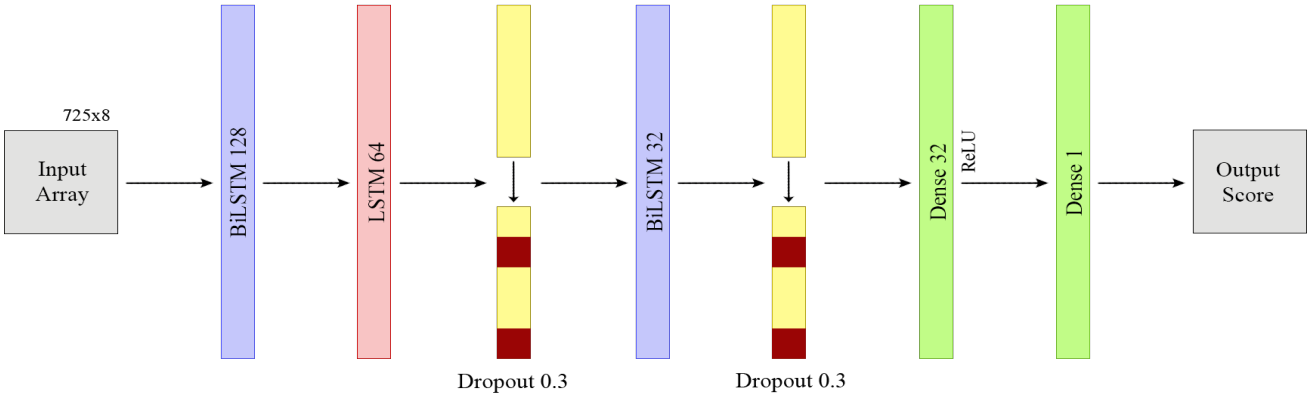

Figure 5 Model Architecture

biomechanically realistic joint trajectories. These two operations combined allow for the generation of varied yet physically plausible motion sequences.

While the augmentation module includes support for temporal shifting (to model delays in motion onset) and joint-space interpolation (to synthesize intermediate poses), these functionalities were disabled in the current experimental

configuration to maintain temporal fidelity.

The entire pipeline was executed programmatically to produce multiple augmented versions of each original sample, each stored in structured CSV format. This resulted in a significantly expanded dataset, enabling robust downstream applications in pose-based skill classification, motion analysis, and karate performance evaluation.

A separate augmented dataset was created for the angles extracted from only the front view to identify the better approach between analyzing from two angles(views) as opposed to one.

3.4. Model Architecture

Our performance evaluation system employs a deep bidirectional Long Short-Term Memory (BiLSTM) network designed for regression of kata performance scores from multivariate time-series data. The architecture is specifically optimized to capture complex temporal dependencies in sequential joint angle trajectories while accommodating variable-length input sequences.

Each kata performance is represented as a time series of joint angles computed from pose landmarks, yielding sequences of shape (T, 8), where T denotes the number of frames (varying across samples) and 8 corresponds to the selected joint angles from key body landmarks including shoulders, elbows, hips, and knees. To enable efficient batch processing, all sequences are zero-padded to a uniform length of 725 frames—the maximum observed sequence length in our dataset.

The complete network comprises approximately 249,409 trainable parameters and is trained end-to-end using supervised learning to minimize the mean squared error between predicted and expert-annotated scores. This architecture effectively captures both forward and backward temporal dependencies while maintaining computational efficiency suitable for real-time deployment scenarios. The model architecture is shown in Figure 5.

3.5. Training Strategy

The model is trained using the Mean Squared Error (MSE) loss function to penalize large deviations from the ground truth kata performance scores. Additionally, Mean Absolute Error (MAE) is tracked as a secondary metric to offer more interpretable insights into average prediction accuracy. Optimization is performed using the Adam optimizer with a learning rate of 0.001, balancing fast convergence with stability.

To ensure generalization and prevent overfitting, an early stopping mechanism monitors validation loss. If no improvement is observed for 10 consecutive epochs, training is halted and the best-performing model weights are restored. Training proceeds for a maximum of 100 epochs with a batch size of 32, allowing the model to efficiently learn both spatial and temporal dependencies in the dual-view pose data.

All training is performed on a machine equipped with an AMD Ryzen 7 7435HS CPU, an NVIDIA RTX 4060 GPU with 8GB of VRAM, and 16GB of GDDR5 RAM. This setup provides sufficient compute power to handle the sequential pose inputs and enables fast, responsive training iterations. The overall strategy is designed to ensure that the model effectively learns the motion patterns and temporal flow necessary for accurate kata skill evaluation.

IV. EXPERIMENTS

We are comparing our scoring approach to the baseline scoring technique in this section.

4.1. Dataset Split

For rigorous model development and fair performance evaluation, the dataset was partitioned following a two-stage protocol. Initially, the complete set of 50 kata sequences was divided into two groups: 37 samples were allocated to the training pool, while the remaining 13 sequences were set aside as the final test set. This test set was strictly isolated

throughout the entirety of model training and validation, ensuring an unbiased measure of generalization.

To address the limited sample size available for training, data augmentation techniques were applied to the 37 training videos, expanding the training pool to 2,400 synthetic sequences. The augmented dataset was then randomly shuffled (with a fixed random seed for reproducibility) and split into training and validation subsets using an 80/20 ratio. Model training was conducted for up to 100 epochs with a batch size of 32, using the validation set to monitor overfitting and tune hyperparameters after each epoch.

Ultimate model selection and all reported performance metrics were based exclusively on the 13 held-out original test sequences. This multi-stage split and strict isolation of the test set ensured that reported results reflect the model's true ability to generalize to unseen, real-world kata performances.

4.2. Baseline Works

To the best of our knowledge, no existing research has been conducted specifically on automated scoring of karate kata performances, particularly using synchronized dual-view video capture systems. This absence of domain-specific literature means our study represents the first attempt at developing a regression-based kata evaluation system that leverages both frontal and lateral camera perspectives for comprehensive movement analysis. Consequently, there is no direct baseline model available for comparison within the karate kata scoring domain, especially for multi-view approaches.

Given this limitation, we reference the Reference Guided Regression (RGR) framework by Jain et. al [26]. For contextual comparison, we consider the RGR(Av) (Expert) variant trained on the Multitask AQA Dataset, which demonstrated the best performance results for complex athletic movements. The fundamental differences in domain, scoring scales, and movement structure prevent direct quantitative comparison. Therefore, our work establishes the first baseline for automated karate kata scoring while drawing broad methodological context from the RGR framework's approach to action quality assessment.

4.3. Evaluation Matrices

For analyzing the accuracy of the regression model, we are using the following metrics: Pearson Correlation, Spearman Rank and MSE.

- *Linear Correlation:* The Pearson correlation coefficient (r) is used to evaluate the strength and direction of the linear relationship between the predicted and ground truth scores. Higher positive values, approaching +1, indicate a stronger linear association and higher model accuracy.
- *Rank Correlation:* The Spearman rank correlation (ρ) is used to measure performance based on the monotonic relationship between the predicted and ground truth scores. It encourages the predicted scores to follow the relative ranking defined by the ground truth, making it robust to non-linearities. Higher values of (ρ) signify a better rank correlation.
- *Mean Squared Error (MSE):* This metric quantifies the average squared difference between the predicted and ground truth scores. It strongly penalizes larger prediction errors and emphasizes that the predicted score should be as close as possible to the ground truth score. Lower values indicate a better fit.

4.4. Ablation Study

To evaluate the contribution of our dual-view capture system, we conducted an ablation study by developing a single-view variant that utilizes only frontal camera footage while maintaining identical architecture and training protocols. The same 37 kata sequences were selected from the original 50-video dataset for training, with identical data augmentation procedures applied to generate 2,400 synthetic samples. The augmented dataset was split using the same 80/20 ratio for training and validation, and evaluation was performed on the same 13 held-out test sequences to ensure fair comparison. The single-view model employs the same bidirectional LSTM architecture, calculating joint angles exclusively from frontal camera perspectives, with all hyperparameters, training procedures, and evaluation metrics remaining consistent between implementations to isolate the impact of viewpoint diversity.

V. RESULTS

The performance of the proposed kata scoring model was rigorously evaluated using standard regression metrics to assess its effectiveness in predicting fine-grained skill levels in martial arts performances. The evaluation methodology prioritized robust and unbiased assessment by conducting all final performance calculations exclusively on a held-out test set comprising 13 original kata sequences from the initial 50-video pool. This test set was completely isolated from the training process and contained no synthetic or augmented samples, ensuring that the reported metrics reflect the model's true generalization capability on authentic, real-world kata performances.

Column Data	Type
6.0	6.351
6.5	6.182
7.0	7.196
7.1	7.204
7.0	7.427
6.5	6.394
6.4	6.358
6.4	6.037
7.2	7.703

Table 3: Actual vs Predicted Scores

While the validation set was required during the training phase for hyperparameter tuning and overfitting prevention, we deliberately refrained from using it for final accuracy calculations as the validation set contained augmented data which could potentially inflate performance metrics and provide a highly optimistic assessment of model capabilities. The scores given by the experts for the test videos against which the scores predicted by the model compared with are given in the Table 3. The model achieved a Mean Squared Error (MSE) of 0.141 indicating minimal average squared

deviation between predicted and expert-annotated scores. In terms of correlation-based metrics, the model achieved a Pearson Correlation Coefficient of 0.784, demonstrating an exceptionally strong linear relationship between predicted and ground-truth scores. Additionally, the Spearman Rank Correlation Coefficient of the model was calculated to be 0.802.

While direct quantitative comparison must be interpreted cautiously due to fundamental differences in domain application, scoring scales, and dataset characteristics, these metrics indicate several advantages of our approach. The superior rank correlation (0.802 vs. 0.76) demonstrates that our model more accurately preserves the relative ordering of performances as assessed by human experts, which is crucial for competitive kata evaluation where precise rankings determine outcomes. More significantly, the dramatic difference in MSE values, though influenced by differing score scales, suggests that our model achieves much finer-grained prediction accuracy within its operational domain.

Model	Spearman Rank Correlation	MSE
Dual-View Approach x	0.802	0.141
RGR(Av) Expert (Jain et al. [26])	0.760	107.25
Single-View Approach	0.434	0.658

Table 4: Performance Comparison of Kata Scoring Approaches

When compared against our single-view variant trained exclusively on frontal footage, the dual-view model achieved substantially superior performance across all metrics: the single-view model attained an MSE of 0.658, Pearson Correlation of 0.425, and Spearman Correlation of 0.434, representing approximately half the correlation performance and nearly five times the prediction error of our final dual-view system. This dramatic performance gap underscores that synchronized orthogonal camera perspectives are essential for capturing the spatial complexity of kata movements. The Table 4 shows how each model compares with each other.

Beyond the quantitative results, these findings represent a transformative advancement in martial arts evaluation. The model's high fidelity in replicating expert judgments enables it to reliably complement human panels, thereby reducing subjectivity and inter-rater variability. By providing continuous, objective scores rather than coarse pass/fail decisions, the system offers practitioners and coaches nuanced, granular feedback on performance progression over time. In competitive settings, the transparency and reproducibility of this scoring pipeline have the potential to enhance fairness and trust in judging outcomes. Furthermore, the model's accuracy can be improved with larger, more diverse datasets. This methodology is also extensible to the scoring of different katas from various karate styles and, with appropriate adaptations, may be applied to other martial arts or broader Action Quality Assessment tasks. Such flexibility underscores the potential of our approach as a generalizable framework for objective performance evaluation across complex human actions.

In summary, the model not only demonstrates state-of-the-art accuracy in replicating expert kata scores but also establishes a practical, deployable framework that promises to revolutionize how martial arts skills are assessed—making high-quality, unbiased evaluation accessible anytime and anywhere.

VI. Future Works

The current study establishes a foundation for automated kata scoring with several promising avenues for enhancement. To create a comprehensive evaluation system, future work could integrate a binary classification module that first categorizes kata performances as "good" or "bad," enabling a two-stage approach where poorly executed kata are efficiently identified and flagged through classification, while competent performances undergo fine-grained regression analysis for precise scoring. For multi-kata evaluation, a kata classification model could be implemented at the pipeline's initial stage to identify the specific kata being performed, subsequently routing the performance to the appropriate kata-specific scoring model. With systematic expansion, this approach could encompass the full spectrum of traditional karate kata across major styles.

REFERENCES

[1] J. Huang, M. Yan, S. Chen, Y. Huang, S. Chen, MagicFight: Personalized Martial Arts Combat Video Generation, Proceedings of the 32nd ACM International Conference on Multimedia (2024) 10833-10842.

[2] L. Hu, Animate anyone: Consistent and controllable image-to-video synthesis for character animation, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 8153-8163.

[3] Z. Husheng, Martial arts moves recognition method based on visual image, Journal of Information Processing Systems, 18(6) (2022) 813-821.

[4] J. Martinez, R. Hossain, J. Romero, J.J. Little, A simple yet effective baseline for 3d human pose estimation, Proceedings of the IEEE international conference on computer vision (2017) 2640-2649.

[5] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, European conference on computer vision (2016) 483-499.

[6] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, Proceedings of the IEEE Conference on computer Vision and Pattern Recognition (2014) 3686-3693.

[7] C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu, Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments, IEEE transactions on pattern analysis and machine intelligence, 36(7) (2013) 1325-1339.

[8] S. Li, S. Huo, W. Ke, Intelligent Decision-Making System for Martial Arts Competition Using Deep Learning, Mobile Information Systems, 2021 (2021) 9920751.

[9] M. Jin, Application and Implementation of Deep Learning for Evaluation of Martial Arts Trainings, Mobile Information Systems, 2022 (2022) 3979817.

[10] E. Wu, H. Koike, Futurepose-mixed reality martial arts training using real-time 3d human pose forecasting with a rgb camera, 2019 IEEE Winter Conference on Applications of Computer Vision (WACV) (2019) 1384-1392.

[11] F.A. Gers, J. Schmidhuber, F. Cummins, Learning to forget: Continual prediction with LSTM, Neural computation, 12(10) (2000) 2451-2471.

[12] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation, 9(8) (1997) 1735-1780.

[13] K. Host, M. Ivašić-Kos, An overview of Human Action Recognition in sports based on Computer Vision, Heliyon, 8(6) (2022).

[14] J.K. Aggarwal, M.S. Ryoo, Human activity analysis: A review, Acm Computing Surveys (Csur), 43(3) (2011) 1-43.

[15] G. Thomas, R. Gade, T.B. Moeslund, P. Carr, A. Hilton, Computer vision for sports: Current applications and research topics, Computer Vision and Image Understanding, 159 (2017) 3-18.
[16] J. Echeverria, O.C. Santos, Toward modeling psychomotor performance in karate combats using computer vision pose estimation, Sensors, 21(24) (2021) 8378.
[17] J. Echeverria, O.C. Santos, KUMITRON: Artificial intelligence system to monitor karate fights that synchronize aerial images with physiological and inertial signals, Companion Proceedings of the 26th International Conference on Intelligent User Interfaces (2021) 37-39.
[18] J. Echeverria, O.C. Santos, KUMITRON: A Multimodal Psychomotor Intelligent Learning System to Provide Personalized Support when Training Karate Combats, (2021) 71-82. https://dl.acm.org/doi/10.1145/3397482.3450730
[19] O.C. Santos, Training the body: The potential of AIED to support personalized motor skills learning, International Journal of Artificial Intelligence in Education, 26(2) (2016) 730-755.
[20] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, 2014 22nd international conference on pattern recognition (2014) 34-39.
[21] K. Zhou, R. Cai, L. Wang, H.P.H. Shum, X. Liang, A comprehensive survey of action quality assessment: Method and benchmark, arXiv preprint arXiv:2412.11149 (2024).
[22] M. Qi, H. Ye, J. Peng, H. Ma, Action Quality Assessment via Hierarchical Pose-guided Multi-stage Contrastive Regression, arXiv preprint arXiv:2501.03674 (2025).
[23] Z. Gao, H. Liu, L. Li, Data augmentation for time-series classification: An extensive empirical study and comprehensive survey, Journal of Artificial Intelligence Research, 83 (2025).
[24] B.K. Iwana, S. Uchida (2021) An empirical survey of data augmentation for time series classification with neural networks. PLoS ONE 16(7): e0254841.
[25] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, H. Xu, Time series data augmentation for deep learning: A survey, arXiv preprint arXiv:2002.12478 (2020).
[26] H. Jain, G. Harit, A. Sharma, Action quality assessment using siamese network-based deep metric learning, IEEE Transactions on Circuits and Systems for Video Technology, 31(6) (2020) 2260-2273.
[27] S. Li, C. Liu, G. Yuan, Martial arts training prediction model based on big data and mems sensors, Scientific Programming, 2021 (2021) 9993916.
[28] H. Zhao, W. Yao, Study on the method of judging standard degree of martial arts movement based on computer vision, 2015 International Conference on Automation, Mechanical Control and Computational Engineering (2015).
[29] P. Parmar, B. Tran Morris, Learning to score olympic events, Proceedings of the IEEE conference on computer vision and pattern recognition workshops (2017) 20-28.
[30] S. Yang, W. Xiao, M. Zhang, S. Guo, J. Zhao, F. Shen, Image data augmentation for deep learning: A survey, arXiv preprint arXiv:2204.08610 (2022).
[31] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of big data, 6(1) (2019) 1-48.
[32] B. Wu, J. Zhou, Video-Based Martial Arts Combat Action Recognition and Position Detection Using Deep Learning, IEEE Access (2024).
[33] M. Walid, M. Ameen, A. Atia, Real-time Detection of Taikyoku Shodan Karate Kata Poses Using Classical Machine Learning and Deep Learning Models, 2023 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC) (2023) 15-20.
[34] M. Daker, F. Elsayaad, A. Atia, Karate kata style classification using pose landmarks and deep learning, 2023 5th novel intelligent and leading emerging sciences conference (NILES) (2023) 228-231.
[35] L.M. Neagu, E. Rigaud, S. Travadel, M. Dascalu, R.V. Rughinis, (2020) Intelligent tutoring systems for psychomotor training--a systematic literature review, in International conference on intelligent tutoring systems, Springer, pp. 335-341.
[36] A. Casas-Ortiz, J. Echeverria, O.C. Santos, (2023) Intelligent systems for psychomotor learning: a systematic review and two cases of study, in Handbook of artificial intelligence in education, Edward Elgar Publishing, pp. 390-421.
[37] H. Pirsiavash, C. Vondrick, A. Torralba, (2014) Assessing the quality of actions, in European conference on computer vision, Springer, pp. 556-571.
[38] J. Xu, Y. Rao, X. Yu, G. Chen, J. Zhou, J. Lu, Finediving: A fine-grained dataset for procedure-aware action quality assessment, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2022) 2949-2958.